## The implications of alternative splicing in the ENCODE protein complement

Michael L. Tress<sup>a,b</sup>, Pier Luigi Martelli<sup>c</sup>, Adam Frankish<sup>d</sup>, Gabrielle A. Reeves<sup>e</sup>, Jan Jaap Wesselink<sup>a</sup>, Corin Yeats<sup>f</sup>, Páll Ísólfur Ólason<sup>g</sup>, Mario Albrecht<sup>h</sup>, Hedi Hegyi<sup>i</sup>, Alejandro Giorgetti<sup>j</sup>, Domenico Raimondo<sup>j</sup>, Julien Lagarde<sup>k</sup>, Roman A. Laskowski<sup>e</sup>, Gonzalo López<sup>a</sup>, Michael I. Sadowski<sup>l</sup>, James D. Watson<sup>e</sup>, Piero Fariselli<sup>c</sup>, Ivan Rossi<sup>c</sup>, Alinda Nagy<sup>i</sup>, Wang Kai<sup>g</sup>, Zenia Størling<sup>g</sup>, Massimiliano Orsini<sup>m</sup>, Yassen Assenov<sup>h</sup>, Hagen Blankenburg<sup>h</sup>, Carola Huthmacher<sup>h</sup>, Fidel Ramírez<sup>h</sup>, Andreas Schlicker<sup>h</sup>, France Denoeud<sup>k</sup>, Phil Jones<sup>e</sup>, Samuel Kerrien<sup>e</sup>, Sandra Orchard<sup>e</sup>, Stylianos E. Antonarakis<sup>n</sup>, Alexandre Reymond<sup>o</sup>, Ewan Birney<sup>e</sup>, Søren Brunak<sup>g</sup>, Rita Casadio<sup>c</sup>, Roderic Guigo<sup>k,p</sup>, Jennifer Harrow<sup>d</sup>, Henning Hermjakob<sup>e</sup>, David T. Jones<sup>l</sup>, Thomas Lengauer<sup>h</sup>, Christine A. Orengo<sup>f</sup>, László Patthy<sup>i</sup>, Janet M. Thornton<sup>e,q</sup>, Anna Tramontano<sup>r</sup>, and Alfonso Valencia<sup>a,r</sup>

aStructural Computational Biology Programme, Spanish National Cancer Research Centre, E-28029 Madrid, Spain; Department of Biology, University of Bologna, 33-40126 Bologna, Italy; HAVANA Group, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; Department of Biochemistry and Molecular Biology and Bioinformatics Unit, University College London, London WC1E 6BT, United Kingdom; Center for Biological Sequence Analysis, BioCentrum-DTU, DK-2800 Lyngby, Denmark; Max Planck Institute for Informatics, 66123 Saarbrücken, Germany; Biological Research Center, Hungarian Academy of Sciences, 1113 Budapest, Hungary; Department of Biochemical Sciences, University of Rome "La Sapienza," 2-00185 Rome, Italy; Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica, E-8003 Barcelona, Spain; Center for Advanced Studies, Research and Development in Sardinia (CRS4), 09010 Pula, Italy; Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland; Center for Integrative Genomics, Genopode building, University of Lausanne, Switzerland; and Pentre de Regulació Genòmica, Universitat Pompeu Fabra, E-08003 Barcelona, Spain

Contributed by Janet M. Thornton, January 31, 2007 (sent for review October 20, 2006)

Alternative premessenger RNA splicing enables genes to generate more than one gene product. Splicing events that occur within protein coding regions have the potential to alter the biological function of the expressed protein and even to create new protein functions. Alternative splicing has been suggested as one explanation for the discrepancy between the number of human genes and functional complexity. Here, we carry out a detailed study of the alternatively spliced gene products annotated in the ENCODE pilot project. We find that alternative splicing in human genes is more frequent than has commonly been suggested, and we demonstrate that many of the potential alternative gene products will have markedly different structure and function from their constitutively spliced counterparts. For the vast majority of these alternative isoforms, little evidence exists to suggest they have a role as functional proteins, and it seems unlikely that the spectrum of conventional enzymatic or structural functions can be substantially extended through alternative splicing.

function | human | isoforms | splice | structure

A lternative mRNA splicing, the generation of a diverse range of mature RNAs, has considerable potential to expand the cellular protein repertoire (1–3), and recent studies have estimated that 40–80% of multiexon human genes can produce differently spliced mRNAs (4, 5). The importance of alternative splicing in processes such as development (6) has long been recognized, and proteins coded by alternatively spliced transcripts have been implicated in a number of cellular pathways (7–9). The extent of alternative splicing in eukaryotic genomes has lead to suggestions that alternative splicing is key to understanding how human complexity can be encoded by so few genes (10).

The pilot project of the Encyclopedia of DNA Elements (ENCODE) (11), which aims to identify all the functional elements in the human genome, has undertaken a comprehensive analysis of 44 selected regions that make up 1% of the human genome. One valuable element of the project has been the detailing of a reference set of manually annotated splice variants by the GENCODE consortium (12). The annotation by the GENCODE consortium is an extension of the manually curated annotation by the Havana team at The Sanger Institute.

Although a full understanding of the functional implications of alternative splicing is still a long way off, the GENCODE set has provided us with the material to make an in-depth assessment of a systematically collected reference set of splice variants.

## Results

Alternative Splicing Frequency. The GENCODE set is made up of 2,608 annotated transcripts for 487 distinct loci. A total of 1,097 transcripts from 434 loci are predicted to be protein coding. There are on average 2.53 protein coding variants per locus; 182 loci have only one variant, whereas one locus, RP1–309K20.2 (CPNE1) has 17 coding variants.

A total of 57.8% of the loci are annotated with alternatively spliced transcripts, although there are differences between target regions chosen manually and those chosen according to the stratified random-sampling strategy (11). The differences stem from gene clusters in the manually selected regions, such as the cluster of 31 loci that code for olfactory receptors in manual pick 9 from chromosome 11 (13). These olfactory receptors are recent in evolutionary origin, have a single large coding exon, and code for a single isoform. This means that although the 0.5% of the human genome that was selected for biological interest has 276 loci, just 52.1% of the loci have multiple variants. In contrast, the regions that were selected in the stratified random-sampling process have fewer loci (158), but 68.7% of the loci have multiple variants (see Fig. 1a). This number is toward the higher end of previous estimates but in line with the most recent reports (14).

Analysis of the data suggests that the GENCODE-validated transcripts are an underestimate of the real numbers of alter-

Author contributions: M.L.T., P.L.M., G.A.R., C.Y., M.A., E.B., S.B., R.C., R.G., J.H., H. Hermjakob, D.T.J., T.L., C.A.O., L.P., J.M.T., A.T., and A.V. designed research; M.L.T., P.L.M., A.F., G.A.R., J.J.W., C.Y., P.İ.Ó., H. Hegyi, A.G., D.R., J.L., R.L., G.L., M.I.S., J.D.W., P.F., I.R., A.N., W.K., Z.S., M.O., Y.A., H.B., C.H., F.R., A.S., F.D., P.J., S.K., and S.O. performed research; M.L.T., P.L.M., A.F., G.R., J.J.W., C.Y., P.İ.Ó., H. Hegyi, M.A., A.G., D.R., J.L., R.L., M.I.S., S.E.A., J.D.W., and A.R. analyzed data; and M.L.T., P.L.M., C.Y., P.İ.Ó., and A.V. wrote the paper.

The authors declare no conflict of interest.

Abbreviation: TMH, transmembrane helices.

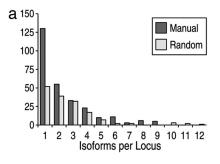
<sup>&</sup>lt;sup>b</sup>To whom correspondence may be addressed at: Spanish National Cancer Research Centres, c. Melchor Fernández Almagro, 3, E-28029 Madrid, Spain. E-mail: mtress@cnio.es.

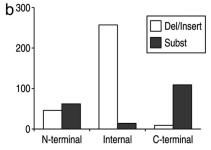
qTo whom correspondence may be addressed. E-mail: helen@ebi.ac.uk.

To whom correspondence may be addressed. E-mail: valencia@cnio.es.

This article contains supporting information online at www.pnas.org/cgi/content/full/ 0700800104/DC1.

<sup>© 2007</sup> by The National Academy of Sciences of the USA





Isoform distribution. (a) The number of isoforms per locus in the manually selected regions compared with the number of isoforms per locus in the regions selected by random stratified procedure. (b) The effect of splicing events on the protein sequence of alternative splice isoforms.

native splicing at the mRNA level. Although the set does include many known isoforms and uncovers numerous previously unrecognized variants, several known variants were not annotated in the initial release. For example, the set annotated just four of the nine experimentally recorded isoforms (15) in locus XX-FW83563B9.3 (TAZ) and only three of six Uniprotrecognized (16) isoforms for locus AC011501.5 (KIR2DL4).

A large proportion of the splice isoforms in the data set have identical protein sequences. These coding sequence-identical variants are alternatively spliced in the 5' and 3' untranslated regions and form an interesting subgroup that may be under independent transcriptional control (17). One locus, AF121781.16 (C21Orf13), has 11 alternative isoforms, all of which are protein sequence-identical. This is not an isolated case: 230 of the 1,097 isoforms are identical, 25 loci have four or more identical isoforms, and 15% of loci with multiple variants code for nothing but protein sequence-identical isoforms.

Splicing Events in Alternatively Spliced Isoforms. We concentrated our analyses on the 214 loci that code for protein sequencedistinct splice isoforms. The results of the analyses can be seen in further detail in the supporting information (SI). We classified the changes brought about by splicing events into six types based on the effect on the protein sequence. Deletions and insertions were pooled because they cannot always be easily distinguished. The results (Fig. 1b) agreed with previous studies (18); internal events are almost always deletions or insertions of single or multiple exons, and splicing events at the C terminus are almost always substitutions.

Although instances of splicing via functionally interesting mutually exclusive exons (2) were very rare in this set, a number of alternative isoforms are generated from translations of different reading frames. For example, alternative splicing between exons 4 and 5 in isoforms 002 and 003 of locus RP1-309I22.1 (TIMP3) leads to a frame shift that causes the fifth exon (corresponding the C terminus of the protein) to be read from a different reading frame. There are only three functionally studied examples of overlapping reading frames in humans. One is INK4a/ARF21 (19), where different transcripts have a coding sequence sharing 3' exons in different reading frames. In this set, 23 separate loci code for variants with overlapping reading frames, suggesting that the frequency of variants with overlapping reading frames might be somewhat higher than previously thought (20).

Few of the variants coded from overlapping reading frames appeared to be functional. We were able to compare nine transcripts where the human and mouse homologue had conserved exonic structure. If the two alternative reading frames evolve under functional constraints, the mutation rate for all three codon positions should be the same, and both frames should have an identical nonsynonymous substitution rate ( $K_a$ ), 21). Only one of the nine pairs of transcripts had identical  $K_a$ values and equal rates of mutation for each coding position.

Signal Peptides and Transmembrane Helices. Of the 1,097 transcripts, 219 were predicted to have signal peptides, accounting for 107 of the 434 loci. Unequivocal loss or gain of signal peptides can be seen in 12 loci, and, in these cases, localization will not be conserved between isoforms. Signal peptide loss/gain is caused by substitution of exons at the N terminus in eight of the loci. This is coherent with earlier findings (18) that showed that most signal peptide gain/loss in alternative splicing comes about through N-terminal exon substitution. One example is the alternative isoform in locus RP1-248E1.1 (MOXD1), which loses 86 N-terminal residues, including the signal peptide, with respect to the principal isoform.

Signal peptide loss in isoform 003 of locus AC010518.2 (LILRA3) is triggered by a 17 residue N-terminal insertion ahead of the signal peptide predicted for the principal sequence. This results in the apparent internalization of the signal peptide and will affect the localization of the protein. If this variant is expressed, there is some evidence to suggest that expression may be disease-associated; the only supporting evidence for this isoform is in the form of ESTs from leukemic blood.

The manually selected regions contain a relatively high proportion of loci that code for proteins with transmembrane helices (TMH). Many of these loci form clearly defined clusters, such as the 31 olfactory receptors on chromosome 11 and the clusters of natural killer cell Ig receptors in manual pick number 1.

Gain or loss of TMH can be observed in 41 loci. In most cases, a single helix is lost relative to the principal splice isoform, but there are also cases where four, five, and even eight membrane sections are missing. Several genes appear to code for both soluble and transmembrane isoforms. For example, the gene UGT1A10 codes for two isoforms of UDP-glucuronosyltransferase 1A10. Isoform 002 has a short substitution in place of the 89-residue C-terminal segment that contains a predicted TMH. All 64 UDP-glucuronosyltransferases deposited in the SwissProt database (16) are annotated as monotopic membrane proteins, and no natural soluble form is known. However, an engineered water-soluble form is reported (22). If expressed at the protein level, isoform 002 would be the first naturally encoded soluble UDP-glucuronosyltransferase.

Splicing events also lead to alternative isoforms in which it is difficult to predict the resulting membrane topology. In locus AC129929.4 (TSPAN32) the principal isoform has four TMH, but the gene also codes for four different splice isoforms that each lose membrane-spanning helices. In isoform 003, the N-terminal helix that acts as both a signal sequence and a membrane anchor (23) is likely to be lost through N-terminal substitution, whereas isoforms 005 and 012 lose the C-terminal TMH. Isoform 014 apparently lacks not just the N-terminal helix, but also the third TMH. This would leave the isoform with two membrane-spanning regions and with the one of the helices oriented in the opposite direction. All these cases must force either a change of structure or polarity.

**Functional Domains.** On occasion, splicing events leave out complete functional domains. This infrequent variation in domain architectures may be biologically meaningful, and these loci are ideal candidates for further study into the potential effects of alternative splicing on function. In this set, the effect was most marked with the immunoglobulin (ig) domain, a functional domain that is overrepresented in the manually chosen regions. For example, isoform 007 from locus AC011501.5 (*KIR2DL4*) is missing the N-terminal ig.

The repeated use of splicing in altering ig fold copy number is of particular interest when attempting to understand the involvement of ig-containing genes in developmental and immune system pathways. Although the numbers of cases is undoubtedly influenced by the significant bias toward ig-like architecture in the manually selected regions, it does suggest that this is not an isolated phenomenon and that it may occur in many other ig-fold-containing proteins. It was also noticeable that no splicing event fell within an ig-domain.

Splicing events occur within Pfam-A (24) hand-curated functional domains in 46.5% of sequence-distinct isoforms, and the figure rises to 71% when all Pfam-defined domains are considered. Although this is a surprisingly high figure, it is still considerably less than might be anticipated. If the same number of splicing events in each alternative isoform were to happen randomly at any of the exon boundaries, they would be expected to fall inside Pfam-A domains in 59.8% of isoforms (84.8% in all Pfam-defined domains). As shown (25), this effect is not due to any correlation between domain and exon boundaries. We found no such correlation.

Although these results do suggest that there is some favorable selection against splicing events that affect functional domains, the proportion of splicing events that occur inside a domains is still high, and, as a result, many of these transcripts are likely to code for proteins with drastically altered structure and function.

**From Gene Expression to Translation.** RT-PCR experiments can confirm mRNA expression, and it is possible to find RT-PCR evidence for many loci. For example, both variants of AF030876.1 (*MEC2P*) have been shown to be expressed (26), and Tsyba *et al.* (27) confirmed the expression of a number of variants from locus AP000303.6 (*ITSN1*).

Although it has been possible to confirm the expression of many alternative transcripts, it is important to know how many of these genes are actually translated into proteins and whether the alternative splice isoforms with the most extreme deletions would become misfolded and quickly removed by the cell degradation machinery. Also, if the proteins are translated and fold properly, what functional role might they play in the cell?

Structures have been resolved for a surprisingly high number of these genes; the Protein Data Bank (PDB) (28) contains structures for proteins from 42 different loci (almost 10% of the total). It was also possible to find homologous PDB structures for more than half of the sequences in the set and, in many cases, to map the changes resulting from splicing events at the protein level onto these structures. We were able to map the sequence of 85 alternative isoforms onto their homologous structures. For 49 of these 85 alternative transcripts, the resulting protein structure is likely to be substantially altered in relation to that of the principal sequence.

A number of alternative isoforms must have radically different structures if they are to fold (see Fig. 2). The gene product of locus RP4-61404.1 (*ITGB4BP*) is 75% sequence identical to a yeast protein with a complex 5-fold  $\alpha$ - $\beta$  propeller structure. Isoform 005 from this locus has an internal substitution of 85 residues (Fig. 2d, marked in purple in the figure), which must disrupt two blades of this very stable structure. In Fig. 3, we demonstrate the effect of large internal deletions or C-terminal substitutions on four splice isoforms from the serpin B cluster in random pick 122. It is clear

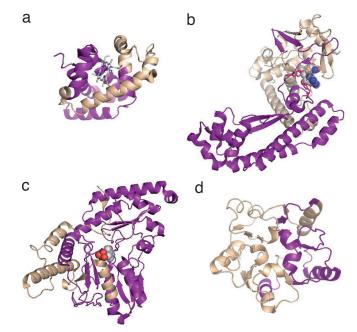


Fig. 2. The potential effect of splicing on protein structure. Four splice isoforms mapped onto the nearest structural templates. Structures are colored in purple where the sequence of the splice isoform is missing. The deletions/ substitutions will mean that the structures of these isoforms would require substantial reorganization from the parent structure. (a) Hemoglobin  $\delta$ -subunit isoform 002 from locus AC104389.18 mapped onto PDB structure 1si4. (b) SET domain-containing protein 3, isoform 002 from locus AL132819.1 mapped onto structure 2h21. (c) Mitochondrial cysteine desulfurase isoform 006 from locus RP1-309K20.1 mapped onto structure 1p3w. (d). Eukaryotic initiation factor 6 isoform 005 from locus RP4-61404.1 mapped onto structure 1g62A. Biologically relevant heteroatoms are shown in space-filling format in b and c, and the heme in a is shown in stick format.

that, in many cases, removing, adding, or replacing part of the protein structure would almost certainly mean that folding and function are severely affected. If these proteins are to fold properly and not aggregate, some alternative structural and functional explanation must be invoked.

Experimental evidence for functional differences between splice isoforms is harder to find. Extensive literature searches for every gene in the ENCODE set turned up just four concrete instances of *in vitro* functional differences between the splice isoforms in this set.

Splice isoform 004 from locus AC034228.1 (*ACSL6*) has a sequence similar internal substitution corresponding to mutually exclusive versions of exon 11 in the transcript. Kinetics assays show that this isoform has conspicuously different ATP-binding affinities (29). A further locus for which there is evidence of distinct functions is XX-FW83563B9.3 (*TAZ*), where it has been shown that alternative isoform 002, which has a 31-residue deletion from skipping the fifth exon, may in fact be the principal isoform, because it is the only isoform to have full cardiolipin metabolic activity (30).

The three experimentally recorded splice variants of locus U52112.3 (*IRAKI*) coincide with the three coding sequence-identical variants in the GENCODE set. IRAK1c (isoform 001) has a 79-residue deletion in relation to the principal sequence (31). IRAK1b (isoform 002) has a 30-residue deletion that results from the use of an alternative 5'-acceptor splice site within exon 12 (32). Both isoforms have been shown to be protein kinase-dead, and neither is phosphorylated by IRAK4. However they do interact with Toll/IL-1 receptor (TIR) signaling factors in the TIR inflammatory cascade, and it has been suggested that the isoforms can function as dominant-negative proteins in TIR-mediated signaling and inflammation.

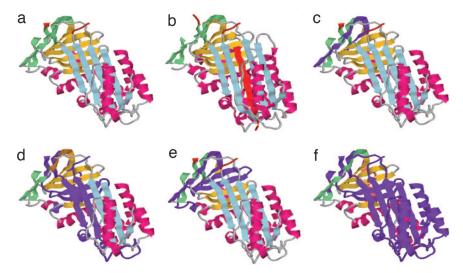


Fig. 3. B serpins. Serpins are protease inhibitors that inactivate their targets after undergoing an irreversible conformational change. (a and b) Serpins exist in an inactivated form (a) that is regarded as being "stressed." Cleavage of the 20-residue RSL region, missing in the structure but with the terminal ends shown in red, causes the RSL region to flip over and fit itself into one of the  $\beta$ -sheets (b, inserted RSL strand in red). This exposes the inhibitory region that inactivates the protease. (c-f) Four splice isoforms from different serpin loci mapped onto the structure of serpinB2 (1  $\times$  7). The sections deleted/substituted in the isoforms are shown in purple. In each case, it appears that splicing is likely to cause the structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantially different fashion. Given that the complex structure to fold in a substantial structure to fold in a substantial structure to fold in the fashion of theof the inhibitor is vital to its unique function, it is not clear why so many apparently deleterious isoforms would be necessary.

Another locus involved in inflammatory response pathways is IL-4, locus AC004039.4, a cytokine thought to play a role in the development of T helper 2 cells. The GENCODE set contains a single alternative splice variant (isoform 002, IL-4d2) that has a deletion of the second exon, a total of 16 residues. Although the presence of isoform IL-4d2 has yet to be demonstrated in the cell, in vitro studies have shown that IL-4d2 retains the ability to bind to IL-4 receptors and acts as a competitive antagonist of IL-4 in monocytes and B cells (33).

The structure of human IL-4 has been well characterized. It is a four-helix bundle with long connecting loops between helices 1 and 2 and 3 and 4 and is held together by three cysteine bridges. The residues coded for by the missing exon coincide with the first of the long loops (see Fig. 4), and to close this loop, a certain amount of structural reorganization relative to the structure of the principal isoform would be necessary. Isoform IL-4d2 has been a favorite target for the homology modeling of splice isoforms, but the size of the gap left by the missing residues and relative inflexibility imposed by the cysteine bridges have hampered predictions. Predicted models have substantially different arrangements of the helices (34–36), and one is even predicted as a knotted structure. Recent results have shown that it is extremely difficult to model even small deletions and insertions with current techniques (37, 38).

Splice Isoforms and Disease. TAZ is not the only locus in the set where doubt has been cast on the biological importance of the principal isoform. At least two other loci (AF030876.1, MECP2) and RP11-247A12.4, PPP2R4) are likely to be annotated with the incorrect principal isoform. It has recently been suggested that, because cDNAs for many genes were cloned from tumor samples, the prevalent isoform may well have been coded from a tumor-specific splice variant rather than the mRNA sequence found in normal tissue (39).

For many alternative variants in this set, the mRNA supporting evidence was found exclusively in cancer cell lines, which suggests that the expression of some of these variants may be associated with disease states. It should be borne in mind, however, that tumor lines are overrepresented in cDNA libraries: 26% of the cDNA libraries annotated in the eVOC pathology annotation (40) are annotated as "normal," whereas 49% are annotated as "tumor."

There has been abundant recent work associating alternative splicing with stresses incurred by cancer and other disorders (41–43), although in many cases the increase in expression of the aberrant variant may be a side effect of the general breakdown of cellular function rather than part of the instigation process. Indeed the importance of alternative splicing in cancer is such that diagnosis can now be carried out by using isoform-sensitive microarrays based on splice isoform profiles (44).

At least two sets of alternative isoforms in this set are implicated in disease. Isoform 011 from locus AC051649.4 (TNNT3) seems to play a role in facioscapulohumeral muscular dystrophy (45), and isoform 006 of locus U52111.6 (*L1CAM*) is involved in CRASH syndrome (46).

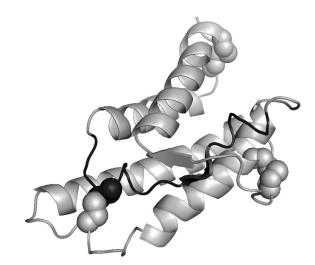


Fig. 4. The difficulty of modeling the structure of isoform IL-4d2. Splice isoform IL-4d2 (isoform 002 from locus AC004039.4) mapped onto structural template 1itl. The section coded for by the missing second exon is colored in dark gray. The cysteine bridges that stabilize the structure are shown in stick format. The missing exon would leave the flanking residues 30 Å apart, and the cysteine bridges mean that the structure has little room for reorganization. Solutions to the modeling problem would have to break the hydrophobic core of the four-helix bundle or break the cysteine bridges by realigning the helices.

## **Conclusions**

This study shows that alternative splicing is commonplace, and the cross-section of alternative splicing events apparent at many different loci points to the possible versatility of alternative splicing in the creation of new functions. However, although alternative splicing has the potential to be an effective way of increasing the variety of protein functions, this still has to be demonstrated at the protein level. Exhaustive literature searches on the genes in this data set unearthed very little evidence of an increase in protein function repertoire.

The effect of splicing on function *in vitro* is known for a few of the alternative isoforms in this set, but even in the cases described above, we are still some way short of knowing their precise role in the cell. Here, detailed and technically complex experimental approaches would be required. At present, most researchers can do little more than hypothesize as to the functional importance of splicing events.

In fact, alternative splicing can lead to a wide range of outcomes, many of which may be undesirable. The large number of alternative splice variants that are likely to code for proteins with dramatic changes in protein structure and function suggest that many of the alternative isoforms are likely to have functions that are potentially deleterious.

The standard path of protein evolution is usually conceived of as stepwise single base-pair mutations. In contrast alternative splicing typically involves large insertions, deletions, or substitutions of segments that may or may not correspond to functional domains, subcellular sorting signals, or transmembrane regions. The deletion and substitution of multiple exons seen in many of these transcripts suggests that splicing is not always a mechanism for delicate and subtle changes and, as a process, may be rather more revolution than evolution.

The substantial rearrangements evident in many of the alternative splice isoforms ought to disrupt their structure and function at the protein level. Changes of this magnitude would normally not be tolerated because of the heavy selection pressure that must oppose such large transformations (47). Unless some external force guides alternative splicing, splicing will lead to as many, if not more, evolutionary dead ends as standard evolutionary paths.

What advantage is there to be gained in the cell from alternative splicing? We cannot rule out the possibility that the expression of alternative transcripts has implications for the control of gene expression, and indeed there are many transcripts that have splicing events outside the translatable re-

- 1. Lopez AJ (1998) Annu Rev Genet 32:279-305.
- 2. Black DL (2000) Cell 103:367-370.
- 3. Modrek B, Lee C (2002) Nat Genet 30:13-19.
- 4. Xu Q, Modrek B, Lee C (2002) Nucleic Acids Res 30:3754-3766.
- 5. Boue S, Letunic I, Bork P (2003) BioEssays 25:1031-1034.
- Wojtowicz WM, Flanagan JJ, Millard SS, Zipursky SL, Clemens JC (2004) Cell 118:619–633.
- Ermak G, Gerasimov G, Troshina K, Jennings T, Robinson L, Ross JS, Figge J (1995) Cancer Res 55:4594–4598.
- 8. Wells CA, Chalk AM, Forrest A, Taylor D, Waddell N, Schroder K, Himes R, Faulkner G, Lo S, Kasukawa T, et al. (2006) Genome Biol 7:R10.
- Matsushita K, Tomonaga T, Shimada H, Shioya A, Higashi M, Matsubara M, Harigaya K, Nomura F, Libutti D, Levens D, et al. (2006) Cancer Res 66:1409–1417.
- 10. Pennisi E (2005) Science 309:80.
- 11. The ENCODE Project Consortium (2004) Science 306:636-640.
- Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, Lagarde J, Gilbert JGR, Storey R, Swarbreck D, et al. (2006) Genome Biol 7:S4.
- Taylor TD, Noguchi H, Totoki Y, Toyoda A, Kuroki Y, Dewar K, Lloyd C, Itoh I, Takeda T, Kim D-W, She X, et al. (2006) Nature 440:497–500.
- Nusbaum C, Zody MC, Borowsky ML, Kamal M, Kodira CN, Taylor TD, Whittaker CA, Chang JL, Cuomo CA, Dewar K, et al. (2005) Nature 437:551– 555.
- Lu B, Kelher MR, Lee DP, Lewin TM, Coleman RA, Choy PC, Hatch GM (2004) Biochem Cell Biol 82:569–576.

gions. It also seems possible that a number of alternative isoforms may have developed a function that is useful for the cell, such as the regulatory role suggested for the IRAK1 and IL-4 isoforms. Despite this, these functional alternative isoforms appear to be the exception rather than the rule.

Cells can encode a great many alternative transcripts; even the conservative estimate from this set suggests that alternative splicing can more than double the number of proteins in the cell. So why are there so many transcripts that appear to encode proteins that are nonfunctional, at least in the classical sense? One answer may be that the organism can withstand these alterations to some extent. It is possible that many potentially deleterious splice variants lie more or less dormant within the gene and are highly expressed only as a result of some disease event. If alternative transcripts in low numbers do not adversely affect the organism, the selection pressure against exon loss or substitution will be reduced, and the new variants will be tolerated, making large evolutionary changes possible.

## **Materials and Methods**

**Data.** The GENCODE annotations used in the paper are available at http://genome.imim.es/biosapiens/gencode/dataset/v2.2.html.

**Databases.** Sequence information is from the Uniprot/Swissprot databases, and structures are from the PDB database. Definitions of protein functional domains, both Pfam-A and Pfam-B domains, come from the Pfam database. The OMIM database (www.ncbi.nlm.nih.gov/omim) was used to relate variants with disease. Relevant literature was found by using iHOP (48).

**Prediction Techniques.** SignalP (49) was used to predict signal peptides, PHOBIUS (50), ENSEMBLE (51), and PRODIV (52) were used to predict TMH; we used BLAST (53) to align sequences and search against the PDB, and exonerate (54) to align human and mouse exonic structure.

Materials and methods are explained in more detail in SI.

We thank Tim Hubbard for support of the project, Christos Ouzounis for initial analysis of the data, and Ana Rojas for technical assistance with the paper. This work was carried out under the umbrella of the BioSapiens Network of Excellence and was supported by European Commission Grant LSHG-CT-2003-503265.

- Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, et al. (2006) Nucleic Acids Res 34:D187– D191.
- 17. Zhang T, Haws P, Wu Q (2004) Genome Res 14:79-89.
- Nakao M, Barrero RA, Mukai Y, Motono C, Suwa M, Nakai K (2005) Nucleic Acids Res 33:2355–2363.
- 19. Quelle DE, Zindy F, Ashmun R, Sherr CJ (1995) Cell 83:993-1000.
- 20. Liang H, Landweber LF (2006) Genome Res 16:190–196.
- Nekrutenko N, Wadhawan S, Goetting-Minesky P, Makova KD (2005) PLoS Genet 1:18.
- Kurkela M, Morsky S, Hirvonen J, Kostiainen R, Finel M (2004) Mol Pharmacol 65:826–831.
- 23. Stipp CS, Kolesnikova TV, Hemler ME (2003) Trends Biochem Sci 28:106-112.
- Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna K, Durban R, et al. (2006) Nucleic Acids Res 34:D247–D251.
- Kriventseva E, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S (2003) Trends Genet 19:124–128.
- 26. Kriaucionis S, Bird A (2004) Nucleic Acids Res 32:1818–1823.
- Tsyba L, Skrypkina I, Rynditch A, Nikolaienko O, Ferenets G, Fortna A, Gardine K (2004) Genomics 84:106–113.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242.
- Van Horn CG, Caviglia JM, Li LO, Wang S, Granger DA, Coleman RA (2005) Biochemistry 44:1635–1642.

- 30. Vaz FM, Houtkooper RH, Valianpour F, Barth PG, Wanders RJA (2003) J Biol Chem 278:43089-43094.
- 31. Rao N, Nguyen S, Ngo K, Fung-Leung W-P, (2005) Mol Cell Biol 25:6521–6532.
- 32. Jensen LE, Whitehead AS (2001) J Biol Chem 276:29037–29044.
- 33. Arinobu Y, Atamas SP, Otsuka T, Niiro H, Yamaoka K, Mitsuyasu H, Niho Y, Hamasaki N, White B, Izuhara K (1999) Cell Immunol 191:161-167.
- 34. Zav'yalov VP, Denesyuk AI, White B, Yurovsky VV, Atamas SP, Korpela T (1997) Immunol Lett 58:149-152.
- 35. Furnham N, Ruffle S, Southan C (2003) Proteins 54:596-608.
- 36. Wen F, Li F, Xia H, Lu X, Zhang X, Li Y (2004) Trends Genet 20:232-236.
- 37. Ginalski K (2006) Curr Op Struct Biol 16:172-177.
- 38. Tress ML, Ezkurdia I, Graña O, López G, Valencia A (2005) Proteins 61:27-45.
- 39. Roy M, Xu Q, Lee C (2005) Nucleic Acids Res 33:5026-5033.
- 40. Law DJ, Labut EM, Adams RD, Merchant JL (2006) Nucleic Acids Res 34:1342-1350.
- 41. Kishore S, Stamm S (2006) Science 311:230-232.
- 42. Ottenheijm CAC, Heunks LMA, Hafmans T, van der Ven PFM, Benoist C, Zhou H, Labeit S, Granzier HL, Dekhuijzen PNR (2006) Am J Respir Crit Care Med 173:527-534.

- 43. Brinkman BM (2004) Clin Biochem 37:584-594.
- 44. Kelso J, Visagie J, Theiler G, Christoffels C, Bardien-Kruger S, Smedley D, Otgaar D, Greyling G, Jongeneel V, McCarthy M, et al. (2003) Genome Res 13:1222-1230.
- 45. Jacob J, Haspel J, Kane-Goldsmith N, Grumet MJ (2002) Neurobiol 51:177-
- 46. Gabellini D, D'Antona G, Moggio M, Prelle A, Zecca C, Adami R, Angeletti B, Ciscato P, Pellegrini MA, Bottinelli R, et al. (2006) Nature 439:973-977.
- 47. Xing Y, Lee C (2006) Nat Rev Genet 7:499-509.
- 48. Hoffmann R, Valencia A (2005) Bioinformatics 21:252-258.
- 49. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) J Mol Biol 16:783-
- 50. Bernsel A, von Heijne G (2005) Prot Sci 14:1723-1728.
- 51. Martelli PL, Fariselli P, Casadio R (2003) Bioinformatics 19:I205-I211.
- 52. Viklund H, Eloffson A (2004) Prot Sci 13:1908-1917.
- 53. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389-3402.
- 54. Slater GS, Birney E (2005) BMC Bioinformatics 6:31.